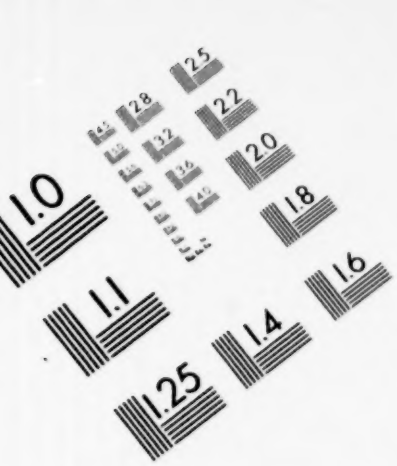
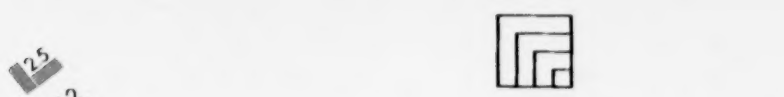
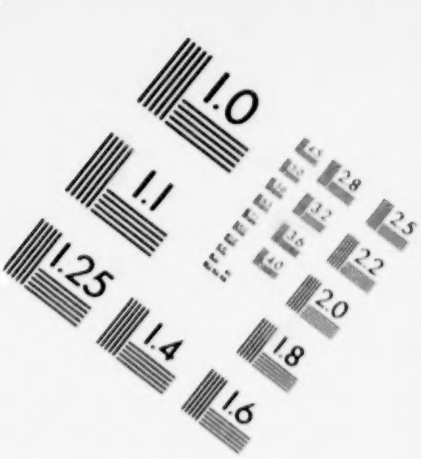
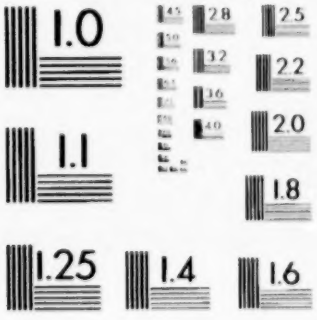


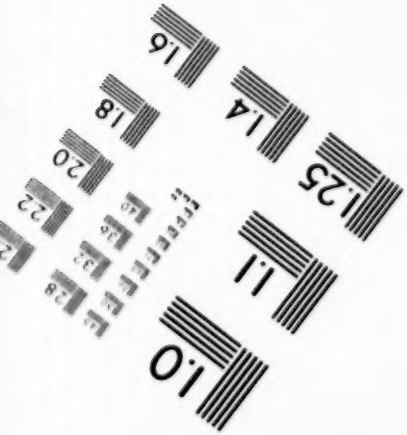


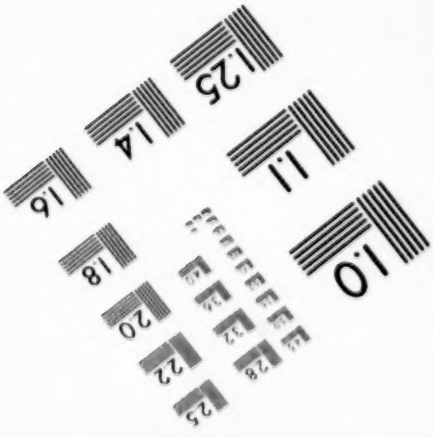
Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202









MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

DOCUMENT RESUME

ED 477 011 IR 021 739

AUTHOR	Fuller, Greg; Simonson, Joe; Tiwari, Ananta; Rebelsky, Samuel A.
TITLE	Clio's Assistants: A Tool Suite for Exploring Student Web Usage.
PUB DATE	2002-06-00
NOTE	7p.; In: ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia & Telecommunications. Proceedings (14th, Denver, Colorado, June 24-29, 2002); see IR 021 687.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	College Students; Computer Software Development; *Data Collection; Educational Resources; Educational Technology; Instructional Materials; Material Development; *Measurement Techniques; Nonprint Media; *Student Behavior; Users (Information) ; *World Wide Web

ABSTRACT

Since the inception of the World Wide Web, faculty members have been developing online course materials. However, there is little careful analysis of how students use these Webs. In particular, do more successful and less successful students use course webs and associated materials differently? Are usage patterns similar to those of printed resources, or do students explore materials differently on the Web? To answer questions like these, educators and researchers need tools that allow them to closely examine student use of Web materials. Building upon user tracking tools that gather information on student usage, including the time each reader arrives at and spends on each page, the use of multiple windows, and the links followed from page to page, the authors implemented Clio's Assistants, a customizable suite of tools that permits exploration of student Web usage patterns both graphically and textually. The graphical tools include simple bar charts, customizable directed graphs, and "replays" of student sessions. Textual tools include: simple statistical summaries, human-readable log files, database queries, and an advanced pattern matching language. Through these tools, one can identify and explore patterns of Web usage. (Contains 11 references.) (Author)

Reproductions supplied by EDRS are the best that can be made

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Clio's Assistants
A Tool Suite for Exploring Student Web Usage

Greg Fuller, Joe Simonson, Ananta Tiwari, and Samuel A. Rebelsky
Glimmer Labs: The Grinnell Laboratory for Interactive Multimedia Experimentation and Research
Grinnell College, Grinnell, Iowa, USA
rebelsky@grinnell.edu

ED 477 011

Abstract: Since the inception of the World Wide Web, faculty members have been developing online course materials. However, there is little careful analysis of how students use these Webs. In particular, do more successful and less successful students use course webs and associated materials differently? Are usage patterns similar to those of printed resources, or do students explore materials differently on the Web? To answer questions like these, educators and researchers need tools that allow them to closely examine student use of Web materials. Building upon user tracking tools that gather information on student usage, including the time each reader arrives at and spends on each page, the use of multiple windows, and the links followed from page to page, we implemented *Clio's Assistants* a customizable suite of tools that permit exploration of student Web usage patterns both graphically and textually. The graphical tools include simple bar charts, customizable directed graphs, and "replays" of student sessions. Textual tools include simple statistical summaries, human-readable log files, database queries, and an advanced pattern matching language. Through these tools, one can identify and explore patterns of Web usage.

1 Introduction

Although a number of tools are used to create educational computing resources, the World Wide Web (Berners-Lee et al. 1994) is perhaps the most popular mechanism for creating computerized educational resources. The Web provides many advantages, including easy design of documents, "universal" access (most Web pages can be accessed from anywhere on the Internet), and the ability to incorporate local and remote documents in a course Web.

Students use course Webs in a variety of ways. For example, some students explore course webs using only one window while others open multiples windows (e.g., one to hold the current problem being studied, another to hold reference materials, and a third to hold current news). Similarly, some students will visit each link on a page while others will be very careful in their selection of links. Are there patterns that successful students seem more likely to use? Do students' usage patterns evolve over time? And, perhaps most importantly, can less successful students benefit from using the patterns of more successful patterns?

Scholars cannot answer any of these questions until there are ways to identify and explore these usage patterns. The goal of Project Clio is to provide tools that allow analysts to identify and experiment with usage patterns. Clio works in three phrases: gathering (Clio's Watchers), synthesis (Clio's Accountants), and analysis (Clio's Assistants). Clio's Watchers are a collection of tools that gather information while a student is browsing the Web. By using the Web Raveler architecture (Kensler and Rebelsky 2000), Clio's Watchers are able to gather information on student usage whether students are on a local or remote site. Because the data gathered by Clio's Watchers are repetitious and make some information (such as time on page) implicit, Clio's Accountants convert the raw data into more useful data that are then stored in a relational database. Finally, Clio's Assistants provide customizable ways to explore those data. The key aspects of Clio's Watchers are that they gather data for a group of individuals (e.g., all members of a course) and that the gather data for all pages those individuals visit, whether they are local or remote.

Clio's Assistants provide the focus of this paper. In Section 2, we describe the primary assistants. In Section 3, we describe a typical interaction with Clio's Assistants. In Section 4, we compare Clio and Clio's Assistants to other hypertext analysis tools. In Section 5, we revisit the need for Clio and Clio's Assistants. Finally, in Section 6, we consider future directions for Clio.

IR021739

BEST COPY AVAILABLE

2 Clio's Assistants: The Suite of Tools

The Clio's Assistants Tool Suite is a collection of tools, both graphical and textual, that allows analysts to explore the ways in which students use the Web. At the core of this exploration is the notion of classification. Although Clio's Assistants permit analysts to explore usage of individual pages (e.g., How long did the average student spend looking at exercise 3?), most assistants allow analysts to consider groups of pages (e.g., When confronted with an exercise, what kind of page did student X most likely visit before entering an answer? Or From what pages were students more likely to leave the local course Web?). The classification of pages is relatively straightforward: For each class of pages, the analyst enters (1) one or more patterns (e.g., "all pages whose URL begins with http://www.cs.grinnell.edu/~rebelsky/Examples/"), (2) a classification (e.g., "Examples") and (3) a shorthand for the classification (e.g., "E"). Classifications are stored permanently in the system. Links may be classified in two ways: by the internal link type (the REL attribute) or by pairs of patterns. Because links depend closely on the particular pages used (e.g., while a link from sect3.5.html to sect2.4.html in the same directory is probably a "prerequisite information" link, not all links between sections are prerequisite links), internal link types are preferred.

Once analysts have classified pages, they may explore the usage logs with both graphical and textual tools. These tools range from simple summaries (textual and bar charts) to complex representations of the data (e.g., as a directed graph or animation over time). Direct access to the "nearly raw" data is also available. In the following paragraphs, we describe the tools in more detail.

Bar Chart: Simple Graphical Summaries Often the best way to begin exploration of data is with a simple overview of the most "popular" parts of a site. Bar charts concisely represent a large amount of data in a simple and quick way. The Bar Chart tool permits analysts to quickly explore a number of comparative relations. For example, bar charts can be generated for number of visits versus URLs (for a group of students) to get a feel for which pages students visit most often, or time spent versus classification to explore how students are dividing their time. We are currently exploring ways to let analysts explore the data within each bar of the chart. For example, upon finding that students spend a lot of time on example pages, an analyst might then ask to see a bar chart of the more popular URLs for example pages.

Summary Statistics: Simple Textual Summaries While the Bar Chart tool provides a general overview of popularity, the overview is limited to one type of thing (URLs, Classification, etc.). The Summary Statistics tool generates a variety of statistical information about the browsing session of a particular user or multiple users. While we are currently exploring the most appropriate information to provide, it currently provides the mean and median time spent on pages, most visited URL, most visited Web domain, most visited classification, and similar data.

Slideshow: Graphical Replays When exploring the patterns of a single student, it is sometimes most useful to watch exactly what the student did: what pages did she visit and in which order, with which on the screen simultaneously. While it would be best to be able to watch over the students shoulder, the Slideshow tool provides a reasonable substitute in that it "replays" the original Web pages visited by the students chronologically, with multiple simultaneous pages shown in different frames. This display is supplemented with additional information, such as time spent on each page and the referring page. This tool is particularly useful when the content of the page, and not just the classification, is of interest.

Directed Graph: Complex Graphical Summaries The tools described so far provide only basic information about usage. How can an analyst find more complex patterns? The Directed Graph tool provides users with a more sophisticated graphical means of examining a browsing history. A directed graph consists of nodes (dots) and directed edges (arrows) that are used to illustrate Web pages and link between those pages, respectively. Different characteristics of the nodes and edges can be set to correspond to different aspects of the pages and links. For example, different colors might represent different classifications and different node sizes might represent different amounts of time spent on a page. Similarly, the color of a link might represent its classification. Alternately, the color of a node might represent the sequential time at which it was visited (making it easier to consider patterns involving multiple windows; similarly colored nodes are likely onscreen at the same time).

Each of a node's four characteristics can be associated with one of eight attributes of the corresponding Web page. The node characteristics are color, size, horizontal position, and textual label. Page attributes include URL, classification, site, title, sequence number (in terms of pages visited in current window), arrival time, maximum time on page, and total time on page. At present, each link has only one customizable characteristic, its color. The color may represent link type, sequence number, or number of times the link was followed.

Each node of a directed graph represents a distinct (window, URL) pair. The directed graph is displayed in one of two frames, with the second frame being reserved for node information. When a node is clicked on, information such as number of page visits, timestamps and URL all become visible in the second frame. Also located in this frame are controls for zooming in and out on the entire graph.

The directed graph may prove beneficial to those who can most easily see patterns with the aid of visuals. Because many nodes and edges can all be examined quickly, patterns of usage become apparent to the user. Combining the graphical means of pattern exploration with the highly customizable attributes of the graph, the directed graph can be an excellent tool for pattern analysis.

Pattern Matching: Complex Exploration of Log Files Once an analyst has identified a pattern through one of the previous tools (e.g., they may see one student who follows many links, but often backtracks, as if they followed the wrong link) or through speculation (e.g., one might postulate that good students always check prerequisite links; the best students quickly realize that they know the prerequisite material and return to the page), analysts may then wish to explore when and how often that pattern occurs within the browsing history of a single student or the class. The Pattern Matching tool is particularly useful for searching through the logs of all students being tracked and letting the analyst see how many and who are following a certain pattern. For example, consider the following sequence of events:

1. The student went to a 'Search Engine' page.
2. The student went to a 'Search Results' page for less than ten seconds.
3. The student went back to a 'Search Engine' page.

This pattern would most likely indicate a failed search. Because the Pattern Matching tool relies on classifications, failed searches can be found at Yahoo and Lycos, not just Google. The Pattern Matching tool has a small language embedded in it that shortens queries to just a few characters. For example, the pattern above would be written in this language as `'SE/SR{<10}/SE'`. SE here stands for Search Engine, SR for Search Results, the '/' indicates, "and then," and the brackets indicate a time interval. The Pattern Matching language provides a number of other special patterns, including an "or", an "at the same time as" and even a way to classify pages on the fly. The results are returned as time intervals for a particular user, and links are provided to the directed graph, the log file, and the slideshow.

We consider the time-on-page pattern particularly useful as it helps one distinguish between apparently similar by actually dissimilar navigation strategies. Consider the pattern "Problem, Reference Page, Example". A student who follows that pattern may be one who wants to read the reference completely and then view some examples. It could also be one who is simply using the reference page as a quick way to get to an example because there is no link to an example from the problem. How can you tell the difference? If the student spends a lot of time on the reference page, the student is probably reading the page. If the student spends only a short amount of time on the reference page, the student is probably using the page as a quick way to get to an example.

"Nearly Raw Data": Log Files and Database Queries As we worked to develop these tools, we interviewed a number of faculty members and instructional multimedia technology specialists to see what features they might want. A few noted, in effect, "While I'd probably use your tools for my initial exploration, I really want access to the raw data for my final analysis". Viewing the log files is another means of analyzing the information available in our database. The log file can be downloaded in CSV format or viewed online. CSV files can be viewed in many spreadsheet programs. Online viewing allows for the creation of links and the ability to link between the various Clio Analysis tools. One can sort and select the data chronologically or by any other information displayed.

Because the data are stored in a relational database, we also expect to provide a mechanism for direct database queries using SQL. However, that remains a future goal.

3 Extended Example

Let us consider a sample session in which one uses Clio's Assistants to explore students Web use. Professor Peabody is interested in studying where and how his students find their information for a particular assignment. Peabody asked his students to turn on Clio while searching for answers on the Web. (For issues of privacy, he first discusses the uses of Clio with the students and obtains their permission.) Several days later, the professor has his students turn in their solutions to their assignments as well as stop using Clio. He then accesses the analysis tools with his favorite Web browser and begins using Clio's Assistants.

Peabody begins by viewing the log files of his students. The information presented to him appears as a series of rows with information about each page the students visited. Having received permission to correlate homework scores with student identifiers, Peabody notes that Student 5 did a wonderful job on his homework and he now views the log file just for Student 5. Because Peabody believes that the log file contains the most information, he begins there. A quick scan of the beginning of the file shows that Student 5 spent a lot of time on the reference pages. However, as Peabody jumps to the end of the file, he sees that most of the entries there are for examples. Two questions quickly arise: Did Student 5's exploration change over time and what kinds of pages did Student 5 use most, the early reference pages or the later example pages?

The second question is perhaps the best place to start. Professor Peabody selects the Bar Chart tool and asks for a comparison of number of visits vs. classification. Both examples and reference pages appear quite frequently. However, Professor Peabody is stunned to see that the most-visited classification was "Course Front Door" and the second most-visited classification was "Search Engine". Two new questions arise: Why did Student 5 revisit the course front door again and again? Was he lost in cyberspace? Perhaps more importantly, why did Student 5 use search engines so frequently?

Peabody asks for summary statistics for Student 5 on the front door. He sees that the average time spent on the page is less then twenty seconds and that the most common referring page is "no referrer". What does that mean? It probably means that Student 5 makes it a habit to jump back to the start of the course web when moving on to new information. Perhaps Student 5 likes to ground his exploration from a comfortable place. A quick check of the directed graph shows a pattern that looks very much like what he postulated: lots of long straight trails that begin with the front door. Is this answer conclusive? No, the best way to find out more is to speak with the student, as in the work described in (Jones and Berger 1996).

Now Professor Peabody wonders what the student was searching for outside of the course web. He sets up a pattern to request all sequences of the form "Search Engine, Results Page, Any Other Page". He expects that by scanning the pages received, he will obtain additional information on what kinds of things Student 5 was looking for. He notices two things. First, that many of the search results that Student 5 selected are within his course web. That suggests that the student is just using the search engine to find pages with the course web. Is that an indication that the links on his pages aren't good, or that Student 5 prefers to search rather than to use links? Another reason to talk to the student. The other thing Professor Peabody notices is that many of the other results are all within a certain domain. It's probably worth exploring whether he should make direct links to those pages.

Before going on to other things, Peabody decides to revisit the question of whether students need a search engine to find pages within his course web. He looks through all students for the pattern "Search Engine, Results Page, Page in My Course Web". He finds that Student 5 is the only student who uses that pattern. Now it's time to consider whether or not that's a technique he should suggest to other students.

4 Conclusions

Most discussions of the World Wide Web in education focus on the direct benefits the Web provides for students: Easier access to material, ability to explore material at the speed and in the manner that best suits the student, availability of a greater variety of materials. Project Clio provides a very different benefit: Clio's goal is to help faculty understand, through data analysis, the pattern and personality of students' usage of the World Wide Web. For Clio to fully assist in the analysis of data, Clio must allow the analyst to explore the various patterns of learning that individuals undergo. Such learning patterns include: active and passive approaches to discovery (Dufresne and Turcotte 1997) and different searching methods (Jones and Berger 1996). By providing a suite of tools that take learning styles into account, we are expanding the boundaries of this underdeveloped area of research. We hope that once scholars are able to identify successful patterns, those patterns can be incorporated into Web sites and taught to less successful students.

The greatest strengths of the Clio's Assistants tool suite are: (1) the diversity of tools available; (2) the interconnectedness of the tools, which permits the results of one tool to be used in another; (3) the ability to record information for a number of students on arbitrary sites, (4) the customizability of the tools, particularly the directed graph tool; (5) the support for information not commonly available, such as time on page and the use of multiple windows; and (6) the richness of the pattern matching language.

5 Future Goals

The first goal of the Clio team is to try the tools in a real classroom situation. We are currently negotiating with some faculty to classify their pages and require their students to use Clio. However, such trials require a more robust version of Web Raveler (currently under development) and a careful consideration of the impact on students. We expect to gather our first data in Fall 2002 and analyze it in the subsequent semester.

Many of the Clio Assistant tools need to be expanded. For example, the Log File tool should be expanded to allow an analyst to select portions of the log file so as to "dig in" to particular parts or patterns. For example, one might first select just the entries for a particular day, then just a particular classification. At the more advanced end of the spectrum, we are hoping to animate the directed graphs so that one can see the graph expand over time, providing an alternative to the slideshow and giving another node characteristic (that is, when the node appears on the screen). We also expect that an animated directed graph will help reveal important aspects of sessions in which students return to certain pages repeatedly.

We are also considering other appropriate tools. As suggested earlier, we are considering ways to give direct SQL access to the data and looking for interesting queries that one might want to make using SQL. We are also considering whether it might be appropriate to present logs in some form of 3-dimensional "world", in which there would be additional opportunities to represent attributes of pages and links (Kmiec et al. 2002). Since the current tools rely on the intuition of those using the tools to find patterns, we are also looking at ways to use data-mining techniques to automatically find patterns (Pinchback et al. 2002).

Finally, we need to consider some subtle issues in more depth. For example, there is a question as to whether a page can receive multiple classifications. For the purpose of the pattern-matching tool, it makes sense to permit a page to be classified in many ways. However, for the directed graph tool, it seems that only one classification can be used at a time (if an example is red and a page about topic X is green, what color should an example about topic X be?). We expect that these kinds of issues will be revealed and clarified as the Clio Assistants are used by a variety of analysts.

References

Becker, R., McLaughlin, K., and Rebelsky. S. A. (1999). Project Clio: Tools for Tracking Student Use of Course Webs. In B. Collis and R. Oliver (Eds.), *Proceedings of the EdMedia 99 World Conference on Educational Multimedia, Hypermedia, and Telecommunications* (Seattle, Washington, June 19-24, 1999), Charlottesville, VA: Association for the Advancement of Computing in Education, pp. 981-986.

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F., and Secret, A. (1994). The World-Wide Web *Communications of the ACM* August 1994 Vol. 37, No. 8, pp. 76-82.

Dufresne, A. and Turcotte, S. (1997). Cognitive Style and Its Implications for Navigation Strategies. In B. d. Boulay, & R. Mizoguchi (Ed.), *Artificial Intelligence in Education: Knowledge and Media in Learning Systems.* Amsterdam: IOS Press, pp. 287-293. URL: http://www.crim.ca/.ipsi/articles/cognitive.pdf. (Accessed May 30, 2001).

Heck, R., Nanjiani, K., Raulerson, E., and Staicut, I. (2000). Clio: Creating History. Unpublished manuscript, August 2000.

Hong, J. and Landay, J A. (2001) WebQuilt: A Framework for Capturing and Visualizing the Web Experience. In *Proceedings of the 10th World Wide Web Conference* (May 1-5, 2001 Hong Kong). URL: http://www10.org/cdrom/papers/602/index.htm (Accessed May 25, 2001; Last modified 15 March 2001).

Jones, T. & Berger, C. (1996) The Pursuit of Knowledge: Interviews and Log Files in Hypermedia Research. In P. Carlson and F. Makedon (Eds.) *Proceedings of the EdMedia'96 Conference on Educational Multimedia and Hypermedia* (Boston, Massachusetts, June 17-23, 1996). Charlottesville, VA: Association for the Advancement of Computing in Education, pp. 342-7.

Jones, T. & Jones, M. (1997) MacSQEAL: A Tool for Exploration of Hypermedia Log File Sequences. In T. Muldner and T. Reeves (Eds.) *Proceedings of the EdMedia'97 Conference on Educational Multimedia and Hypermedia* (Calgary, Canada, June 14-19, 1997), 1997. Vol. 1. Charlottesville, VA: Association for the Advancement of Computing in Education, pp. 539-44.

Kensler, A. and Rebelsky, S. A. (2000). Web Raveler: An Infrastructure for Transforming Hypermedia. In J. Bourdeau and S. Heller (Eds.) *Proceedings of the EdMedia 2000 World Conference on Educational Multimedia, Hypermedia, and Telecommunications* (June 26-July 1, 2000, Montreal, Quebec, Canada), pp. 479-484. Charlottesville, VA: AACE, 2000.

Kmiec, A., Koomjian, J., Lindsey, D., and Rebelsky, S. A. (2002). Clio's Worlds: Three-Dimensional Visualizations of Student Web Usage. Poster abstract accepted to EdMedia 2002 World Conference on Educational Multimedia and Hypermedia (June 24-29, Denver, Colorado).

Pinchback, M. P., Stoltzfus, R., Tiwari, A., and Rebelsky, S. A. (2002). Clio's Intuition: Using Data Mining to Find Patterns of Student Web Usage. Poster abstract accepted to EdMedia 2002 World Conference on Educational Multimedia and Hypermedia (June 24-29, Denver, Colorado).

Raulerson, E., and Staicut, I. (2000) Clio: A Course Web History Analyzer. Public presentation, Pew Midstates Science and Mathematics Consortium Undergraduate Research Symposium in Mathematics and the Physical Sciences (St. Louis, MO, November 5, 2000).

Acknowledgements

We are particularly grateful to the Grinnell College Mentored Advanced Project (MAP) fund for the funding for the research described within this paper. The work benefited from the modern computing equipment in the department of Mathematics and Computer Science, equipment funded by Grinnell College and The Arthur Vining Davis foundation. Our work builds upon the prior work of others in the Grinnell Laboratory for Interactive Multimedia Experimentation and Research (Glimmer), particularly Andrew Kensler, who built the first version of Web Raveler, the Summer 2000 Web Raveler team (Peter Broadwell, Elizabeth Hulse, Chris Kern, and Erin Nichols), who developed a more robust and modular implementation, the Summer 1998 Clio Team (Raphen Becker and Kevin McLaughlin), who built the first prototype of the Clio tools, and the Summer 2000 Clio Team (Rachel Heck, Kumail Nanjiani, Eleanor Raulerson, and Izabela Staicut), who built the Web Raveler plugins to gather and synthesize data. We are also grateful to the other Summer 2001 research students who provided guidance in the Web Raveler system and who gave us suggestions for our own work: the Annotations team (Khong Lovan, Elias Vafiadis, and Joshua Vickery) and the Links team (Angela Kmiec and Melissa Piper Pinchback). Finally, we appreciate the time of the Grinnell College faculty and staff who took the time to discuss the analysis of student data with us.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").